Mycobacterium tuberculosis complex genetic diversity

Brudey, K; Driscoll, Jeffrey; Rigouts, Leen; Prodinger, Wolfgang; Gori, Andrea; Al-Hajoj, Sahal; Allix, Caroline ; Aristimuno, Liselotte; Arora, Jyoti; Baumanis, Viesturs; Binder, Lothar; Cafrune, Patricia; Cataldi, Angel; Cheung, Soonfatt; Diel, Roland; Ellermeier, Christopher; Evans, James T; Fauville-Dufaux, Maryse; Ferdinand, Severine; Garcia de Viedma, Dario

DOI:
10.1186/1471-2180-6-23

BMC Microbiology

Research article

Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology

Karine Brudey[1], Jeffrey R Driscoll[2], Leen Rigouts[3], Wolfgang M Prodinger[4], Andrea Gori[5], Sahal A Al-Hajoj[6], , Caroline Allix[7], Liselotte Aristimuño[8], Jyoti Arora[9], Viesturs Baumanis[10], Lothar Binder[11], Patricia Cafrune[12], Angel Cataldi[13], Soonfatt Cheong[14], Roland Diel[15], Christopher Ellermeier[16], Jason T Evans[17], Maryse Fauville-Dufaux[7], Séverine Ferdinand[1], Dario Garcia de Viedma[18], Carlo Garzelli[19], Lidia Gazzola[5], Harrison M Gomes[20], M Cristina Guttierez[21], Peter M Hawkey[17], Paul D van Helden[22], Gurujaj V Kadival[23], Barry N Kreiswirth[24], Kristin Kremer[25], Milan Kubin[26], Savita P Kulkarni[23], Benjamin Liens[1], Troels Lillebaek[27], Ho Minh Ly[28], Carlos Martin[29], Christian Martin[30], Igor Mokrousov[31], Olga Narvskaïa[31], Yun Fong Ngeow[14], Ludmilla Naumann[32], Stefan Niemann[33], Ida Parwati[34], Zeaur Rahim[35], Voahangy Rasolofo-Razanamparany[36], Tiana Rasolonavalona[36], M Lucia Rossetti[12], Sabine Rüsch-Gerdes[33], Anna Sajduda[37], Sofia Samper[38], Igor G Shemyakin[39], Urvashi B Singh[9], Akos Somoskovi[40], Robin A Skuce[41], Dick van Soolingen[25], Elisabeth M Streicher[22], Philip N Suffys[20], Enrico Tortoli[42], Tatjana Tracevska[10], Véronique Vincent[21], Tommie C Victor[22], Robin M Warren[22], Sook Fan Yap[14], Khadiza Zaman[35], Françoise Portaels[3], Nalin Rastogi*[1] and Christophe Sola*[1]

Address: [1]Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Guadeloupe, [2]Wadsworth Center, New York State Dept. of Health, Albany, NY, USA, [3]Mycobacteriology Unit, Prince Leopold Institute of Tropical Medicine, Antwerp, Belgium, [4]Dept. Hygiene Microbiology and Social Medicine, Innsbruck Medical University, Innsbruck, Austria, [5]Dept of Infectious Diseases, Institut of Infectious Diseases, Milano, Italy, [6]Department of Comparative Medicine, King Faisal specialist Hospital and Research Center, Riyadh, Saudi Arabia, [7]Laboratoire de la Tuberculose, Institut Pasteur de Bruxelles, Belgique, [8]Universidad Centrooccidental Lisandro Alvarado, Barquisimeto, Venezuela and Universidad de Zaragoza, Spain, [9]All India Institute of Medical Sciences, New Delhi, India, [10]Biomedical Research and Study Center, Riga, Latvia, [11]Institut for Hygiene, Microbiologie and Tropical Medicine, Austria, [12]Universidade Federal do Rio Grande de Soul, Brazil, [13]Instituto de Biotecnologia INTA, Castelar, Argentina, [14]Dept of Medical Microbiology and Pathology, faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia, School of Public Health, [15]University of Düsseldorf, Heinrich-Heine-University, Düsseldorf, [16]Dept of Internal Medicine II, University of Regensbourg, Germany, [17]Public Health Laboratory, Hearltlands Hospital, Birmingham, UK, [18]Dept of Clinical Microbiology and Infectious Diseases, Hospital Gregorio Marañon, Madrid, Spain, [19]Dept. of Experimental Pathology, Medical Biotechnology, Infection and Epidemiology, Pisa University, Pisa, Italy, [20]Laboratory of Molecular Biology applied to Mycobacteria, Dept. Mycobacteriosis, Oswaldo Cruz Institute, Rio de Janeiro, Brazil, [21]Centre National de Référence des Mycobactéries, Institut Pasteur, Paris, France, [22]MRC Centre for Molecular and Cellular Biology, Dept of medical Biochemistry, University of Stellenbosch, Tygerberg, South Africa, [23]Laboratory Nuclear Medicine Section, Isotope group, Bhabha Atomic Research Centre c/T.M.H. Annexe, Parel, Mumbai-400012, India, [24]Public Health Research Institute, Newark, NJ, USA, [25]Mycobacteria reference unit, Diagnostic Laboratory for Infectious Diseases and Perinatal Screening, National Institute of Public Health and the Environment, Bilthoven, The Netherlands, [26]Municipal Institute of Hygiene, Prague, Czech Republic, [27]Statens Serum Institute, Int. Ref. lab. for Mycobacteriology, Copenhagen Denmark, [28]Institute of Hygiene and Epidemiology, Hanoi, Vietnam, [29]Universidad de Zaragoza, Zaragoza, Spain, [30]Laboratoire de Bactério-virologie-hygiène, CHU Dupuytren, Limoges, France, [31]Institut Pasteur de Saint-Petersbourg, Saint Petersbourg, Russia, [32]Bavarian Health and Food Safety Authority, Oberschleissheim, Germany, [33]Forschungszentrum, National Reference Center for Mycobacteria, Borstel, Germany, [34]Dept of Clinical Pathology, Padjadjaran University, Dr. Hasan Sadikin Hospital, Bandung, Indonesia, [35]Tuberculosis Laboratory, International Centre for Diarrhoeal Research, Dhaka, Bangladesh, [36]Institut Pasteur de Madagascar, Tananarive, Madagascar, [37]Dept of Genetics of Microorganisms, University of Lódz, Lodz, Poland, [38]Servicio Microbiología, Hospital Universitario Miguel Servet, Zaragoza, Spain, [39]State Research Center for Applied Microbiology, Obolensk, Russian Federation, [40]Dept. of Respiratory Medicine School of Medicine Semmelweis University, Budapest, Hungary, [41]Veterinary Sciences Division, Department of agriculture for Northern Ireland, Belfast, UK and [42]Centro regionale di Riferimento per i Micobatteri, Laboratorio de Microbiologia e Virologia, Ospedale Careggi, Firenze, Italy

Published: 06 March 2006

*BMC Microbiology*2006, **6**:23 doi:10.1186/1471-2180-6-23

Received: 08 November 2005
Accepted: 06 March 2006

This article is available from: http://www.biomedcentral.com/1471-2180/6/23

Abstract

Background: The Direct Repeat locus of the *Mycobacterium tuberculosis* complex (MTC) is a member of the CRISPR (Clustered regularly interspaced short palindromic repeats) sequences family. Spoligotyping is the widely used PCR-based reverse-hybridization blotting technique that assays the genetic diversity of this locus and is useful both for clinical laboratory, molecular epidemiology, evolutionary and population genetics. It is easy, robust, cheap, and produces highly diverse portable numerical results, as the result of the combination of (1) Unique Events Polymorphism (UEP) (2) Insertion-Sequence-mediated genetic recombination. Genetic convergence, although rare, was also previously demonstrated. Three previous international spoligotype databases had partly revealed the global and local geographical structures of MTC bacilli populations, however, there was a need for the release of a new, more representative and extended, international spoligotyping database.

Results: The fourth international spoligotyping database, SpolDB4, describes 1939 shared-types (STs) representative of a total of 39,295 strains from 122 countries, which are tentatively classified into 62 clades/lineages using a mixed expert-based and bioinformatical approach. The SpolDB4 update adds 26 new potentially phylogeographically-specific MTC genotype families. It provides a clearer picture of the current MTC genomes diversity as well as on the relationships between the genetic attributes investigated (spoligotypes) and the infra-species classification and evolutionary history of the species. Indeed, an independent Naïve-Bayes mixture-model analysis has validated main of the previous supervised SpolDB3 classification results, confirming the usefulness of both supervised and unsupervised models as an approach to understand MTC population structure. Updated results on the epidemiological status of spoligotypes, as well as genetic prevalence maps on six main lineages are also shown. Our results suggests the existence of fine geographical genetic clines within MTC populations, that could mirror the passed and present *Homo sapiens sapiens* demographical and mycobacterial co-evolutionary history whose structure could be further reconstructed and modelled, thereby providing a large-scale conceptual framework of the global TB Epidemiologic Network.

Conclusion: Our results broaden the knowledge of the global phylogeography of the MTC complex. SpolDB4 should be a very useful tool to better define the identity of a given MTC clinical isolate, and to better analyze the links between its current spreading and previous evolutionary history. The building and mining of extended MTC polymorphic genetic databases is in progress.

Background

Each year, 9 million new cases of tuberculosis (TB) are recorded, of which 2 million result in fatality. Diagnostics, chemotherapy and vaccination are available, however, the disease is far from being eradicated [1]. Many genetic loci within the *Mycobacterium tuberculosis* complex (MTC) genomes are polymorphic and may be used for molecular evolutionary studies [2]. Among these, the Direct Repeat locus (DR), which consists of alternating identical DRs and variable spacers can be assessed using the "Spoligotyping" fingerprinting method thousands of different patterns [3]. DR loci are members of a universal family of sequences, designated as CRISPR [4], whose physiological role is poorly known [5,6]. Spoligotyping was previously shown to be useful for both clinical management and molecular epidemiology of MTC [7]. When used in association with variable-number of DNA tandem-repeat (VNTR) [8] or Mycobacterial-interspersed-repetitive-units (MIRU) [9], spoligotyping is a fast, robust, and cost effective genotyping technique, alternative to traditional IS*6110*-RFLP fingerprinting. These methods are also designated as MLVA (Multiple-locus variable number tandem repeats analysis) [10]. Since 1999, we have built and released genetic diversity databases of the MTC DR locus as an attempt to analyze MTC population structure, and to assess the complexity of global TB transmission and of the

underlying spatial and temporal evolution of the TB genetic landscape. Indeed, previous studies have shown that the host's geographical origin is predictive of the clinical isolate of tuberculosis being carried, since there is an apparent stable association of TB bacilli populations with their human hosts in various environments [11], hence a strong phylogeographical clustering of TB bacilli population. We hypothesize that co-evolution between human beings and bacilli, and vertical transmission (in the household), must have been the main mode of tuberculosis transmission throughout centuries and even millenniums [12].

MTC organisms were also shown to evolve clonally [13]. Hence, the reconstruction of the population structure of this species may be an indirect way of assessing its main host's (*Homo sapiens sapiens*) migratory and demographic history [14]. Indeed, tuberculosis may have affected early hominids and it is tempting to speculate that the MTC originated in East-Africa [15]. Its expansion to the rest of the world may have coincided with the waves of human migration out of Africa, with potential back migration from Asia to Sub-Saharan Africa [16]. If the past phylogeny of MTC is likely to have involved horizontal gene transfer events, however, these events are no longer observed [17]. All these clues suggests the pioneer roles of geography, demography and human migration history in shaping today's MTC population structure [11]. Consequently, the current concepts of "natural evolving communities" or "clusters of bacilli", are important: (1) for TB epidemiology – as the global pandemia should be considered as a network of outbreaks of more or less circumscribed clones [18,19] – (2) for molecular ecology, evolutionary and population genetics and phylogeography, – as today's MTC genomes are likely to include cryptic information on their passed and present history in their changing environments [20,21] – and last but not least (3) for systematics and infra-species taxonomy [22]. The two first databases were poorly representative of the worldwide MTC diversity [23,24], whereas the third update was already more representative [25,26]. Although these studies did not allow for a definitive rebuilding of the MTC limb and twig history of the TB tree, their use combined to bioinformatical data-mining methods, allowed to previously classify MTC in eight to ten main genetic lineages [27], a classification that has now been validated using SNPs [28,29].

In this new study, we data-mined an updated SpolDB4 version, which contains 1939 STs representing a total of 39,295 clinical isolates originating from 122 countries. Considering the known diversity of the origin of patients, which was documented in some cases, SpolDB4 is representative of a total of 141 countries. The SpolDB4 update adds 1126 newly defined STs and provides a higher resolution picture of the worldwide MTC genome diversity assessed by spoligotyping. However, the new challenge is to link the genetic variability of MTC with the clinical variability of TB, whatever the setting, whether in high burden countries or in densely populated areas such as India or China.

Results

Classification of spoligotypes

Classification of SpolDB4 spoligotype patterns into meaningful lineages

The listing of ST alleles with their distribution by country of location and a presumed sub lineage/lineage label is provided in additional file 1. 51 new countries are represented in SpolDB4. Out of 39,295 spoligotypes patterns, n = 35,925 are found in 1939 STs (91.4% of the isolates,) and 3370 (8.6%) are orphan patterns, totalling 5,309 individual alleles. Two approaches, a statistical, and a mixed expert-based/bioinformatical one were used to data-mine SpolDB4 to classify spoligotypes. Results are summarized in Figure 1.

Results obtained by the statistical approach

The 20 most frequent STs totalled 17,701 isolates (49.3% of the clustered isolates). The 50 most frequent STs increased clustering to 61.8% (n = 22,219). These 50 most frequent types are shown in figure 1. Three types did not receive a lineage label, ST46, ST51 and ST210. ST46 and ST51 are patterns prone to genetic convergence, similarly as reported for ST4, whose ancestors can either be ST33 or ST34 [30]. ST210 (also designated as HN24, a Principal Genetic Group (PGG) 3 strain) was first described in a study done in Texas and is almost restricted to the USA [31]. The other 47 most prevalent spoligotypes belong to known genetic lineages or are defined variants.

Results obtained by the mixed expert-based/bioinformatical approach

Figure 1 also describes a total of 62 remarkable lineages/sub lineages. This classification was obtained with the use of a dedicated software that search for similarities between patterns (SpolNet, P. Abdoul *et al*. unpublished, see material and methods section). Since SpolDB4 is a mixed *M. tuberculosis* (human) and *M. bovis* (human or bovine) isolates database, 237 STs were found to belong to the *M. bovis* subspecies (n = 5710), whereas 1702 STs (n = 33,585) were not *M. bovis*. The calculation of the genetic diversity index "H" -defined as $H = 2n\ (1 - \Sigma x_i^2)/2n-1$ where n = number of individuals and x_i is the frequency of the i^{th} allele-gives a value of 0,98. H only slightly improved (+0.6%) compared to SpolDB3. This shows that the exponential increase of data was not reflected by the increase in bacterial diversity description, as most data were already known (over fitting phenomenon, a limitation of this study). Indeed, a quantitatively updated

ST n°	lineage/sublineage	Binary spoligotype	Octal**	Total	Frequency
181	AFRI1		770777777777671	270	0.75
331	AFRI2		774077607777071	74	20
438	AFRI3		774077777777071	30	0.08
1	Beijing		000000000003771	4051	11.28
482	BOV1		676773777777600	2401	6.68
683	BOV2		664073777777600	2770	7.71
479	BOV3		640013777777600	204	0.57
647	BOV4		200003777377600	166	0.46
691	BOV1-variant 1		676573777777600	175*	0.48*
665	BOV1- variant2		616773777777600	120*	0.33*
676	BOV2-variant1		664073774037600	177*	0.49*
679	BOV2-variant2		664073777763600	428*	1.19*
592	canettii		000000000101000	9	0.03
26	CAS1-Delhi		703777740003771	1076	3
21	CAS1-Kili		703377400001771	158	0.44
25	CAS1-variant		703777740003171	128*	0.35*
288	CAS2		700377740003771	79	0.22
405	Zero		777777000000011	19	0.05
100	MANU1		777777777773771	75	0.21
54	MANU2		777777777763771	160	0.45
1378	MANU3		777777777770771	28	0.08
539	microti		000000000000600	38	0.11
593	PINI1		074000037777600	11	0.03
637	PINI2		000000007777600	93	0.26
236	EAI-5		777777777413771	1014	2.82
48	EAI1-SOM		777777777413731	450	1.25
19	EAI2-Manilla		677777477413771	544	1.51
89	EAI2-Nonthaburi		674000003413771	72	0.2
11	EAI3-IND		477777777413071	432	1.2
139	EAI4-VNM		777777774413771	347	0.97
591	EAI6-BGD1		777777757413771	169	0.47
1898	EAI7-BGD2		777777770003771	68	0.19
109	EAI8-MDG		477777377413771	63	0.17
47	H1		777777774020771	1236	3.44
62	H1-variant1		777777774020731	171*	0.44*
2	H2		000000004020771	296	0.82
50	H3		777777777720771	1963	5.46
127	H4		577777777420771	374	1.04
20	LAM01		677777607760771	484	1.35
17	LAM02		677737607760771	411	1.14
33	LAM03		776177607760771	509	1.42
60	LAM04		777777607760731	148	0.41
93	LAM05		777737607760771	264	0.73
64	LAM06		777777607560771	172	0.48
41	LAM07-TUR		777777404760771	160	0.45
290	LAM08		777777606760771	22	0.06
42	LAM09		777777607760771	1930	5.37
61	LAM10-CAM		777777743760771	432	1.2
59	LAM11-ZWE		777777606060771	240	0.67
209	LAM12-Madrid1		776017607760771	24	0.07
53	T1		777777777760771	6412	17.85
280	T1-RUS2		770000777760771	63	0.18
34	S		776377777760771	648	1.8
52	T2		777777777760731	840	2.34
135	T2-Uganda		777777777760730	48	0.13
37	T3		777737777760771	277	0.77
149	T3-ETH		777000377760771	221	0.62
40	T4		777777377760771	168	0.47
39	T4-CEU1		777777347760471	133	0.37
44	T5		777777757760771	298	0.83
58	T5-Madrid2		777777557760771	103	0.29
254	T5-RUS1		777760007760771	100	0.28
1737	Tuscany		777760017760771	11	0.03
451	H37Rv		777777477760771	147	0.41
627	T3-OSA		741737777760771	19	0.05
119	X1		777776777760771	956	2.66
137	X2		777776777760601	1069	2.98
38	X2-variant1		777776407760601	125*	0.34*
92	X3		700076777760771	699	1.95
91	X3-variant1		700036777760771	143*	0.39*
197	X3-variant2		700036777760731	244*	0.67*
51	undesignated		777777777760700	158*	0.43*
4	LAM3/S		000000007760771	118*	0.32*
73	T3-variant		777737777760731	114*	0.32*
210	undesignated		776037777760771	104*	0.29*
49	H3-variant		777777777720731	102*	0.28*
46	undesignated		777777770000000	95*	0.26*

Figure 1

Bioinformatical (62 lineages/sub lineages prototype patterns) and statistical (50 most frequent) classification analysis of SpolDB4. First column ST n°: Shared-type (ST) number of prototype pattern for the linage/sub lineage. Second column: lineage/sub lineage name. Third column: Binary spoligo display with black-white squares for respectively hybridizing-non-hybridizing spacers. Fourth column: Octal code (in red: defining octal rule). Fifth column: total absolute number of isolates of the subclass when variant ST spoligos are included (using SpolNet). Sixth column: same but expressed as percentage of total clustered isolates. * Total number and Frequency for these types are already included in their mother clade if known. Undesignated types are counted within the T1-ill-defined lineage. ** in red: octal rule defining the genotype.

SpolDB3 (on 817 STs) would cluster 31,292 isolates, whereas the new 1122 STs aggregate only 4633 additional clustered isolates. Further improvement of H would require new data entries, possibly from yet unrepresented settings and/or high TB-burden countries such as India and China.

New genetic lineages within M. tuberculosis *complex*

SpolDB4 defines 62 genetic lineages/sub lineages (figure 1). *M. bovis* strains were divided in 3 sub lineages corresponding to ST prototypes 482, 683 and 479. A new signature for *M. microti* is suggested (presence of spacers sp37-38, ST539) [32], instead of ST641 in SpolDB3. *M. caprae* and 2 sub lineages of *M. pinipedii*, a new member of MTC [33], were added. Within the *M. africanum* subspecies, more is known today on its taxonomical status, thanks to improvement of spoligotyping through the 68 spacers format and thanks to the discovery of other lineage-specific genetic markers [34,35]. Using a dedicated software (*structure* version 2, [36,37]) to infer the population structure of the *M. africanum* spoligotyping dataset of SpolDB4, one would suggest the existence of at least 4 populations in SpolDB4 (results not shown), however more data on the genetic diversity of *M. africanum* will be required to be able to get a clearer picture of the global population structure of this pathogen.

New genetic lineages within M. tuberculosis *stricto sensu*

Among *M. tuberculosis* stricto sensu, new visual rules defining 22 lineages/sub lineages are described. The previously defined Central-Asian (CAS) lineage was split into CAS1-Delhi type (ST26) found mainly in India and in the Indian subcontinent [38,39], and CAS1-Kilimanjaro (ST21) found in Tanzania [40]. Within the East-African-Indian (EAI) lineage, new prototypic spoligotyping-signatures for 4 sub lineages are presented (EAI2-Nonthaburi, EAI6-Bangladesh/1, EAI7-Bangladesh/2 and EAI8-Madagascar). Douglas *et al.* designated the EAI2 clade as the "Manila family" [41]. We further linked the Nonthaburi group of strains from Thailand [42] to this lineage (results not shown). EAI3 and EAI4, are now being shown as phylogeographically specific from India and Vietnam respectively, with suggested designations of EAI3-IND and EAI4-VNM. Two new lineages from Bangladesh are found, designated as EAI6-Bangladesh/1 (58.1% of isolates from Bangladesh) and EAI7-Bangladesh/2 (91.2% of isolates from Bangladesh). EAI6-BGD1 harbours specificity for the eastern part of the South Asian region since it is also found in neighbouring Myanmar (results not shown).

Within the Haarlem (H) lineage, a 4th sub lineage (H4) is tentatively added. It is characterized by the absence of spacers 29–31 and 33–36 (prototypes ST127 and/or ST777). More than 60% of ST127 isolates are localized in Armenia, Austria, Finland, Georgia, Iran, and Russia. A likely related pattern (ST777) is found in Saudi Arabia. An hypothesis is that these strains could represent an intermediate genetic link between the previously defined Haarlem-1 (H1, ST 47) and Haarlem-3 (H3, ST50) genotypes. Type ST777, which shows a single spacer difference from ST 127, is also found in Kazakhstan, Russia, and Georgia (n = 26). These isolates are likely to be identical to the recently described "Ural" family of strains [43]. They suggest the prominent role of Central Asia as a hub of migratory routes of *Homo sapiens sapiens* and its role in the history of infectious diseases.

Within the Latin-American-Mediterranean (LAM) lineage, we rebaptised the LAM7 sub lineage as LAM7-Turkey since recent results suggested that ST41 is predominant in Asia Minor [44]. Similarly, the LAM10 sub lineage was renamed as LAM10-Cameroun [45,46]. Two sub lineages are new, LAM11-ZWE (ST 59) with 57.8% of isolates originating from Zimbabwe [47], and LAM12-Madrid1 (ST209) [48]. The LAM11-ZWE is likely to be identical to the recently described Meru family found in Tanzania [40]. The "Manu" family, a new family from India, which could be an ancestral clone of principal genetic group 1 strains [39], is tentatively sub-divided into Manu1 (deletion of spacer 34), Manu2 (deletion of spacers 33–34), and Manu3 (deletion of spacers 34–36). The central role of India, and more generally of Asia in tuberculosis evolutionary history is more and more evident. The S lineage which is highly prevalent in Sicily and Sardinia, could be identical to the F28 clade in South Africa [30,49]; the existence of this genotype family was confirmed in SpolDB4, however its origin remains unknown. The "T" families (modern TB strains) stayed ill-defined with more than 600 unclassified STs. They were stratified into 5 subclades (T1-T5) based on single-spacer differences. 8 nested clades, with robust spoligotyping-signatures were extracted; with the exception of "Tuscany", their names were built using their proximate upper-clade designation (T1 to T5), followed by their presumed geographical specificity: T3-Ethiopia (ST149); T5-Russia/1 (ST254), T1-Russia/2 (ST280), T3-Osaka (ST627), T5-Madrid/2 (ST58), T4-Central Europe/1 (ST39), T2-Uganda (ST135), and "Tuscany" (ST1737). ST149 was previously shown to be frequent in Ethiopia and in Denmark among Ethiopian immigrants. This low-banding IS*6110* clone had been identified based on IS*6110*-RFLP as early as 1995 and represented 36.2% of isolates from this country [50]. ST254 and ST280 were repeatedly isolated from clinical isolates in Russia, in former Russian soviet republics and in Northern and Eastern European countries (Estonia, Finland, Georgia, Latvia, Poland, Russia). ST1737, with a single spacer difference from ST254, was recently found in Italy and designated as "Tuscany" [51]. ST627 was identified for the first time in Finland and repeatedly found in the Okayama district and elsewhere in Japan [52]. T5-Madrid/

2 isolates were previously found to be characteristic of Spanish-related settings [48]. T2-Uganda, which was first described by Niemann *et al.*, was repeatedly found in East Africa, and at least 7 STs with a prototypic signature are linked to East-Africa. Last but not least, the T4-CE1 (for Central-Europe/1) was tentatively identified based on similarity between ST39 found in Europe and South Africa, and some likely derived genotypes found both in South and North America (ST94, ST430, ST1258). Among those, ST1258 represents the most prevalent spoligotype detected by the Inuit's community in Nunavik, Canada. Whether this type has been introduced by a casual European contact into this community, or has been endemic among the Inuits is currently under investigation [53].

The X genotype family (X1-X3 sub lineages) was initially described thanks to data-mining [27]. This family is today a well-characterized IS*6110* low-banding family, duly characterized by IS*6110*-RFLP as well as by high-throughput genetic methods [54-57]; it is prevalent in UK, in USA and in former British colonies. Lastly, a Vietnamese genotype family characterized by the absence of IS*6110* insertion elements, was shown to bear a specific spoligotyping signature which is characterized by the deletion of spacers 19–41 (ST 405), and was designated as the "Zero copy" clade [58].

Comparison between the supervised and an unsupervised Naïve Bayes statistical approach of classification of spoligotypes

Data-mining and clustering techniques are the focus of intense research in Information and Bio Sciences [59]. Classification of spoligotypes, given the almost infinite dimension of theoretical allelic number -n = 2^{43} in the current format that does not detect the complete set of known spacers-, is not a trivial task. A recent attempt to use a statistical approach of classification of spoligotypes through the utilization of a Naïve Bayes algorithm and a mixture model was suggested [60]. A good correlation was found between the two techniques for a large number of genotype families defined, with strong support of the stability value. However, the mixture model also suggested the existence of new spoligotype signatures and a total of 48 families (N1 to N48) [60]. Some of these signatures are confirmed by this study, others are not. A similar approach using a Markov model is currently under development in our laboratory.

Unclassified spoligotypes

454 STs (23.4%) did not correspond to any pattern recognition rule when data were mined automatically using SpolNet. Nonetheless, we attributed a family label to these spoligotypes too by extending the rules defined initially, and subsequently by a visual assessment on each unclassified ST. Thus, 314 more STs were tentatively classified with a family label, and only 131 STs (7% of the total) remained unclassified. These STs harboured patterns with either important blocks of deletion, disseminated deletions, mixed signatures, or patterns, which at this time did not correspond to any known spoligotype-signatures.

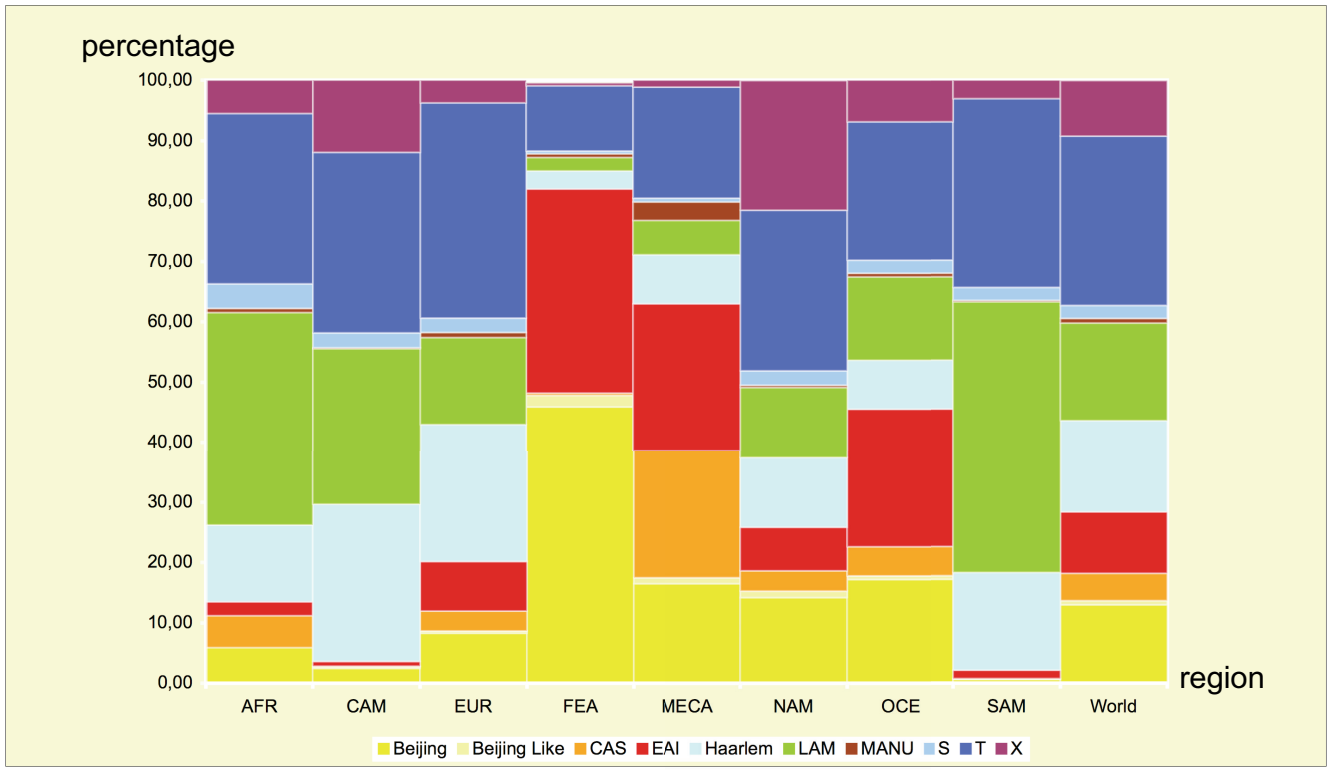
Results

Population genetics

The worldwide distribution of data points was assessed within eight regions. Isolates from Europe and North America represented 65.5% of all entries. Africa, Far-East Asia, Middle East and Central Asia, and South America were equally represented (6.5% to 8.3% of entries) whereas Central America and Oceania were underrepresented (3.6% and 1.1% of entries). The overrepresentation of *M. bovis* from Europe and South America (about 30 and 25% respectively) limits the interpretation on global *M. bovis* genetic diversity, but also reflects the reality of the importance of beef cattle economy in South America (Brazil and Argentina) and Europe. Orphan spoligotypes, represented 35.7% and 16.9% of the isolates in Europe and North America respectively, and ranged from 1.2% to 11.5% in the other regions.

Figure 2 is a synthetic histogram of the distribution of 10 main lineages in the studied continents. In brief, Beijing and Beijing-like strains represent about 50% of the strains in Far East-Asia and 13% of isolates globally. In Europe, the Haarlem lineage represents about 25% of the isolates. In South America, about 50% of the strains belong to the LAM family. Three major genotypic families (Haarlem, LAM, and T) are the most frequent in Africa, Central America, Europe and South America. Outside Europe, The Haarlem strains were mainly found in Central America and Caribbean (about 25%), suggesting a link of Haarlem to the post-Columbus European colonization [61] (Figure 3). The presence of the LAM family is highest in Venezuela (65%) [62], in the Mediterranean basin (e.g. 34% in Algeria, 55% in Morocco, 30% in Spain), and in the Caribbean region (30% in Cuba and Haiti, 17.4% in French Guiana, 15% in French Caribbean islands) (Figure 3A). The "ill-defined" T genetic family, was found in all continents, and corresponded to about 30% of all entries in the database. Undoubtedly, MLVA and/or SNPs data will improve the knowledge of the identity of these isolates designated as "T" by default [26].

The Beijing family of strains is prevalent in Far-East-Asia, but also in Middle-East-Central Asia and Oceania (45.9%, 16.5% and 17.2% respectively) (Figure 3). The Beijing genotype which may have been endemic in China for a long time [63] is emerging in some parts of the world, especially in countries of the former Soviet Union, and to a lesser extent in the Western world [64]. The East-African-Indian (EAI) family is also highly prevalent in these areas



Abbreviations AFR = Africa, CAM = Central America, EUR = Europe, FEA = Far-East Asia, MECA = Middle-East and Central Asia, NAM = North America, OCE = Oceania, SAM = South America

Figure 2
Percentage of main spoligotyping-defined MTC genotype families within SpolDB4 (Beijing, Beijing-like, CAS, EAI, Haarlem, LAM, Manu, X, T), by studied continents and worldwide. Abbreviations : AFR = Africa, CAM = Central America, EUR = Europe, FEA = Far-East Asia, MECA = Middle-East and Central Asia, NAM = North America, OCE = Oceania, SAM = South America.

(33.8% in Far-East-Asia, 24.3% in the Middle East and Central Asia, and 22.9% in Oceania). The EAI lineage is more prevalent in South-East Asia, particularly in the Philippines (73%; [41]), in Myanmar and Malaysia (53% ; [65]), in Vietnam and Thailand (32% ; [62]) (Figure 3).

The CAS1-Delhi family is essentially localized in the Middle-East and Central Asia, more specifically in South-Asia, (21.2%), and preferentially in India (75%; [38,66]). It is also found in other countries of this region such as Iran, and Pakistan [67,68]. It has also been found in several others regions (Africa, 5.3%; Central America, 0.1%; Europe, 3.3%; Far-East-Asia, 0.4%; North-America, 3.3%; Oceania, 4.8%). In Europe and Australia, these strains were frequently found to be linked with immigrants from South Asia [24].

Lastly, the X family is highly prevalent in North America (21.5%) and Central American (11.9%) regions. It could be linked to an Anglo-Saxon ancestry, as it has been encountered in English-colonized areas such as in the United Kingdom, United States, Australia, South Africa and in the Caribbean [27]. However, according to other investigators, this group of strains is currently correlated with African-Americans, a fact however that may not represent the ancestry of this genotype [69]. More studies should be done to clarify this issue.

Analysis of the spreading and epidemiological status of MTC clones
We further analyzed the epidemic status of each spoligotyping-defined clone as shown by the index C1 and C2 [26] as well as the inter-continental match between STs. Results are shown in Table 1 and Table 2. Briefly, the Spreading Index (SI) represents a mean number of occurrences of a clone, independent of the setting. Fourteen types were defined as "epidemic" (SI>25), 65 as "common"(10<SI<25), 669 as "recurrent" (SI<10) and 1090 as "rare" (SI ≤ 2). The Table 1 shows the results based on combination of C1 and C2 which provides 12 classes ranging from the highly localized (endemic) and numerous (epidemic), to the highly spread (ubiquitous) and infrequent (rare) genotypes. When analyzed geographi-

Table 1: A. Definition of the variables (MC, Ar, SI) used in SpolDB4 to define : (1) the geographic index C1 (Endemic, Localized, Ubiquitous) (2) the quantitative index C2 (Epidemic, Common, Recurrent, Rare). B. Distribution of the 1939 Shared-types in 12 classes.

Name	Abbreviation	Type of data, Def.	Rules for Definition of Qualifiers (C1 and C2)
A Definitions			
Matching Code	MC	1–8 digits, built by linking region codes	If 1 digit, then C1 = Endemic (genotype found in one macroregion only) If 2 digits, then C1 = Localized (genotype found in two macroregions) If ≥ 3 digits, go to Area section below for further interpretation
Area	Ar	numeric, n° of countries in which a given SIT is found	If MC ≥ 3 digits and Areas ≤ 5; C1 = Localized If MC ≥ 3 digits and Areas ≥ 6; C1 = Ubiquitous (genotype found in more than three macroregions)
Spreading Index	SI	numeric, mean indicator of spreading independent of geography SI = n/Ar where n is the absolute value for a given shared-type	If SI ≥ 25; C2 = Epidemic If 10 < SI < 24; C2 = Common If 3 < SI < 9; C2 = Recurrent If SI ≤ 2; C2 = Rare
B Distribution of the 1939 genotypes in 12 class			
Endemic Epidemic	6	Endemic Recurrent	286
Localized Epidemic	2	Localized Recurrent	207
Ubiquitous Epidemic	6	Ubiquitous Recurrent	176
Endemic Common	31	Endemic Rare	501
Localized Common	14	Localized Rare	587
Ubiquitous Common	20	Ubiquitous Rare	82
NA*	21	Total	1939
*not applicable			

cally using the Matching Code (MC), a total of 824 types are found within a single macro region ("Endemic", 42,5%), and 564 types are present in exactly two settings ("localized", 29%). Types present in three macro regions but found in five or less areas are also defined as localized and totalled 246 types. The Intercontinental match of these types was not analyzed further. All other types, being present in at least three continents and in at least six areas, or present in four or more continents, (n = 551), were declared as "ubiquitous". The Table 2 shows the number of endemic types per continent and presents the matching results between "localized" types. Endemic types are likely to represent local clones, current end points of evolution, either because of extinction (for ancient clones) or because epidemiological transmission was not yet followed by mutation for emerging ones. Independently of recruitment, the number of endemic types is minimal in Oceania and Central America (0.005 endemic type/occurrence), two regions that have experienced a negative migratory balance for centuries. In all other continents, with a 0.015 to 0.03 endemic type/occurrence, and slightly more in Middle-East and Central Asia (0.037), rates of endemism appear to be similar, a feature resulting from a combination of the intrinsic molecular clock of the DR locus, the age of the tubercle bacilli, and the historical tuberculosis transmission waves.

Toward an interactive system of geographic atlas of genotype frequency data of Mycobacterium tuberculosis

Figure 3 and 4 illustrate a mixed representation (by absolute case number and by frequency) of the distribution of the most frequent STs shown in Figure 1, and grouped by genetic lineage, for the following six MTC lineages: Beijing, *M. bovis*, Central-Asia, East-African-Indian, Haarlem and Latin-American-Mediterranean. These Figures provides the best display of the global phylogeographical structure of the MTC population. Similar results focused in Europe (data not shown) and using SpolDB4 suggests the existence of fine geographic genetic clines between four prevalent genotypes belonging to the modern MTC types, i.e. ST53-T1, ST50-Haarlem3, ST47-Haarlem1, ST42-LAM9 [70].

Discussion

In this study, we data-mined an updated international spoligotype database of the *M. tuberculosis* complex, SpolDB4, both for improving classification of MTC genomes, and for presenting a more reliable snapshot picture of the global and local population genetics of tubercle bacilli. Considering the known diversity of the origin of patients, SpolDB4 represents clinical isolates from a total of 141 countries. This is to our knowledge the largest collaborative effort to describe the worldwide genetic structure of MTC.

Table 2: Analysis of the geographical Inter and Intra-continental matches between the shared-types found within 1 or 2 settings (endemic types n = 824) and localized types (n = 564), within and between the 8 macro-regions. The geographical analysis of the matches between localized types found in 3 macroregions and 5 settings or more (n = 246) and between the ubiquitous types (found in more than 3 regions (n = 284) was not done. Number of intra and intercontinental matches between STs detected between the 8 macro regions previously defined.

Macroregion (code)	Africa	Americas			Europe	Asia		Oceania
		North America	Central America/ Caribbean	South America		Middle-East and Central-Asia	Far-East Asia	
Africa (1)	73	29	2	7	79	7	3	0
North America (2)		138	8	35	131	34	45	4
Central America (3)			7	0	9	1	2	0
South America (4)				94	69	ND	0	3
Europe (5)					351	56	25	3
Middle-East Asia (6)						99	8	0
Far-East Asia (7)							61	4
Oceania (8)								1

ND = not done

The scaling-up that represents SpolDB4 relatively to SpolDB3 (4×), allowed new sub lineages to be discovered. However, it also showed the limit of the approach of using spoligotyping only to define the precise identity of a given MTC clone, since over fitting was observed. Combined DR, MLVA, SNPs, Region of Differences polymorphic datasets are now required to improve our knowledge of MTC genomes diversity.

Today's observed pattern of phylogeographical diversity of MTC is undoubtedly the result of both a deep ecological differentiation and of a more recent demographic and epidemic history. It is tempting to speculate, especially since the publication of the recent studies done on *M. canettii* [15,71], that TB is as old as humanity. It is also tempting to hypothesize that the EAI ancestral strains spread back from Asia to Africa through India concomitantly to human migrations [72], and that evolution gave rise to the CAS lineage, and possibly to all "modern" TB lineages. Recently, Mokrousov *et al.* used the Beijing lineage as a model to compare its phylogeography with human demography and Y chromosome-based phylogeography [73]. Further work using other genetic markers will help to better define the retrospective demographical history of the various sub lineages described here.

Such an endeavour as the building of SpolDB4, and more generally the building of a representative genetic diversity database, should however assume limitations concerning both the quality and representativity of data. We partially eliminated the first problems by carefully double checking many datasets visually, reinterpreting other datasets by asking for the autoradiography results sent as electronic files, or by simply excluding datasets harbouring systematic genotyping errors. The procedure of examining only STs improves data quality by minimizing artefacts analysis. Some internationally agreed upon recommendations to improve the quality of spoligotyping are also in progress and will be published elsewhere. The second problem (representativity) is more chronic and difficult to solve. More financial means should be devoted to improving both mycobacteriology diagnostic and genotyping facilities in countries where the outbreaks prevail. Another limitation using spoligotyping is the study of mixed populations of bacilli [74]. Indeed, MLVA is appropriate, contrary to spoligotyping, to reveal the existence of dual infections, an issue which, especially in high prevalence countries, was probably underestimated and that may sometimes jeopardize spoligotyping results. Future studies should attempt to evaluate whether admixture models can in fact explain some yet undefined mixed spoligotyping signatures.

One of the major lessons from this collaborative effort is that new markers such as MLVA, SNPs or others are needed to improve our knowledge on the population structure of MTC. MLVA provides an improved knowledge of MTC clonal complexes, a strategy for the MTC which could even be more effective than the Multi-Locus-Sequence-Typing (MLST) approach [55,75]. MLVA allows the investigator to limit the quantity of required genotyping to only type epidemiologically or phylogenetically informative markers, depending on the branch depth (regarding time) to which a given data set should to be analyzed [76]. Rapid changes in MTC genotyping methodologies have lead to numerous ongoing debates about the choice of the best genotyping strategy [77-79].

Our efforts to characterize the classification of the tuberculosis strains present within populations have also led to

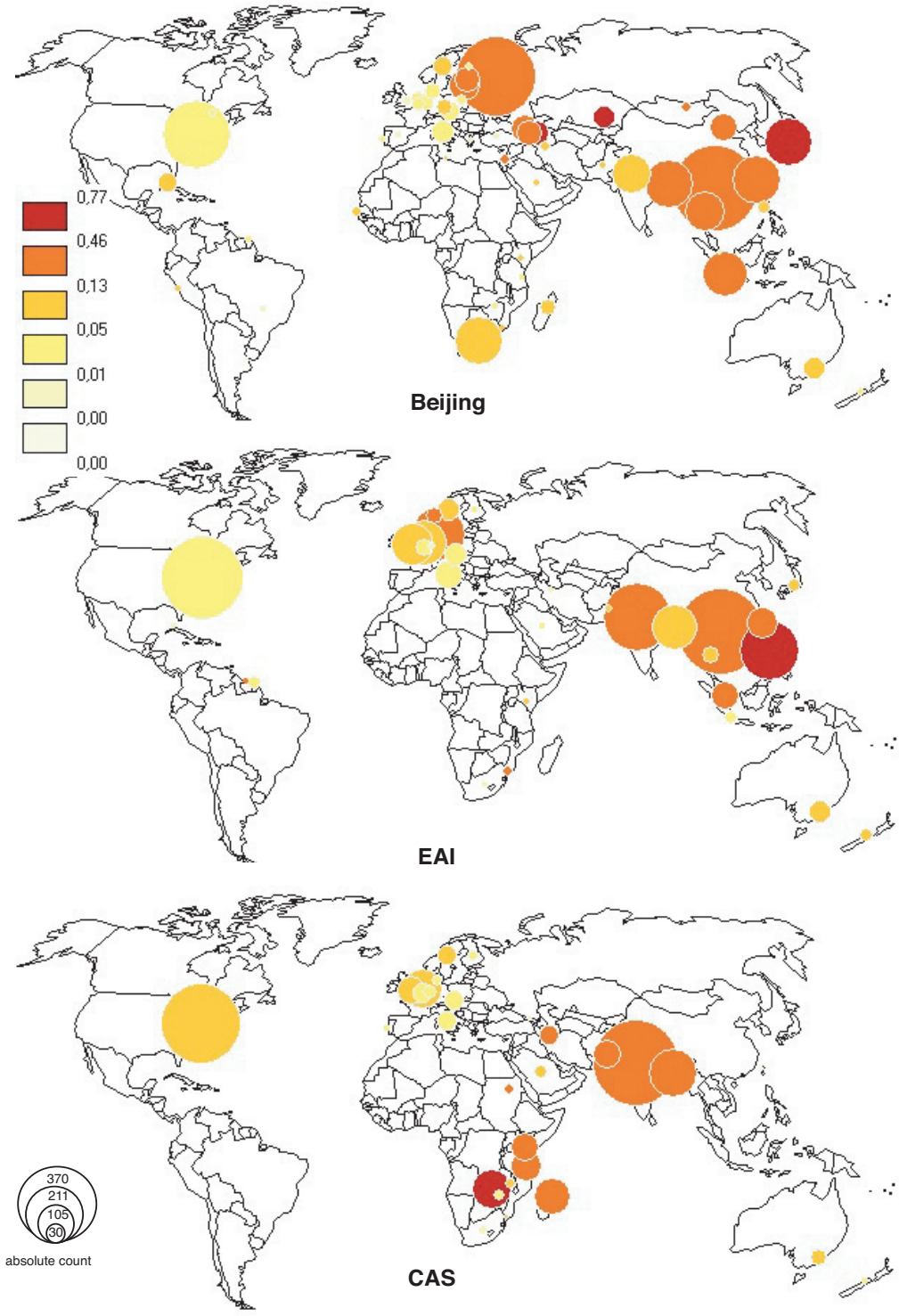


Figure 3
Synthesizing World Maps showing absolute (diameter) and percentage (colour) numbers of 3 genotype families within each country: Beijing; EAI (East-African Indian) CAS (Central Asia). These maps were built on an updated SpolDB4 on 2005 September 14th, on clusters of the 50 most frequent shared types as shown in Table 1, for a total of n = 17212 isolates (Beijing n = 4042, EAI n = 1684, CAS n = 1022).

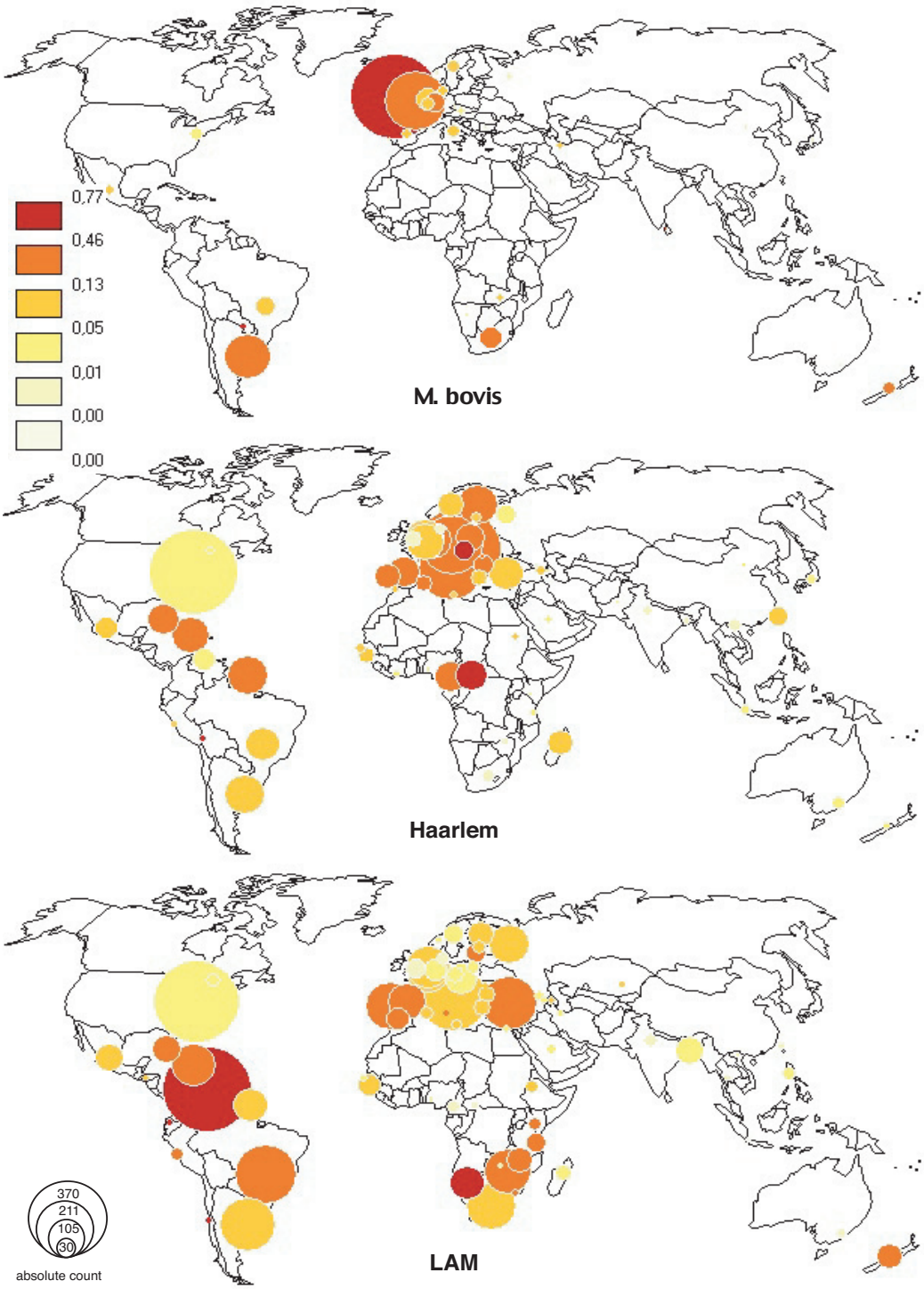


Figure 4
Synthesizing World Maps showing absolute (diameter) and percentage (colour) numbers of 3 genotype families within each country: *M. bovis*; Haarlem; Latin-American and Mediterranean (LAM). These maps were built on an updated SpolDB4 on 2005 September 14th, on clusters of the 50 most frequent shared types as shown in figure 1, for a total of n = 17212 isolates (*M. bovis* n = 3888, LAM n = 3400, Haarlem n = 3176). Maps were built using Philcarto (P. Waniez, version 4.38).

an increased understanding of their global distribution. Our results do not fully agree with recently published results [11] suggesting that *M. tuberculosis* did not possess a finer geographic structure than the one defined within broad regions (East Asia, Africa, Europe, Philippines and Americas). On the contrary, our results demonstrate that the genetic diversity of *M. tuberculosis* genomes and hence their population structure, is strongly linked to geography at a fine geographical scale, thus reinforcing the importance of localized effort to control tuberculosis and to consider the global tuberculosis pandemia as the sum of very different and genetically separate individual outbreaks. Future work should focus on an adequate modelization of the database according to demography and tuberculosis prevalence, to present a more realistic quantitative TB genetic landscape.

Linking these results for a clinical benefit for individual patients remains another challenge. In particular, we need to better understand why in certain areas a small number of strains are causing a disproportionate number of cases of the disease and we need to better understand the effects of host's genetic and environmental variability on the presentation of the disease [80]. In the Western world TB occurs mainly due to reactivation of disease in the older age class and immigration in younger ones. Consequently the characteristics of the bacterial population are those of an old outbreak under extinction, with the superimposition of new characteristics due to cases of importation and recent transmission. The more homogeneous population structure in high-burden countries is more likely to reflect the ongoing transmission in all age classes. Selection of particular genotypes (e.g. due to vaccination with *M. bovis* BCG) and clonal selection [81] has also been suggested [50,82]. Previous colonization histories, and of course more deeply rooted anthropological structures as well as geographic isolation, may have also contributed to the complex tuberculosis genetic landscape.

Conclusion

The SpolDB4 database is by far one of the largest publicly available database on *M. tuberculosis* complex genetic polymorphism with a universal nomenclature system of spoligotyping (octal or arbitrary ST number) as well as a global epidemiological information system. Our results suggest the existence of fine geographical genetic clines that may correlate to the passed *Homo sapiens sapiens* demographical history. Further combined (multimarkers) genetic databases, as well as local finer scale analysis are in progress to further analyze the complexity of the evolutionary history of the MTC.

Methods

Database

A total of 39,295 entries were collected in an Access® database. The Previous STs numbered 1 to 817 were previously described in SpolDB3. ST633, 714, 729, 770, 797, which were missing in the former version were replaced by new genotypes. SpolDB4 is available [see additional file 1] and can also be downloaded (sorted/unsorted versions) at: [83]. A dedicated interactive website version is available [93].

The distribution by continent and country of isolation is shown below. Macro region name, defining acronym, identification number and total number of patterns (into brackets) are respectively: Africa AFR -1 (n = 3121), Central-America CAM -2 (n = 1353), Europe EUR -3 (n = 13624), Far-East Asia FEA -4 (n = 2624), Middle East and Central Asia MECA -5 (n = 2639), North-America NAM -6 (n = 9153), Oceania OCE -7 (n = 235), South-America SAM -8 (n = 3176).

Origin of spoligotypes

The spoligotypes were either obtained at the Pasteur Institute of Guadeloupe using the 43-spacers format and home-made membranes [84,85], received from co-investigators and collaborating laboratories, or retrieved from local molecular epidemiological studies or diversity-driven published articles. The sampling is still far from being representative for all countries, and in many cases, limited epidemiological and patient information on the isolates were available. However, we assume that this " convenience sample " is representative of world-wide TB genotypes and that, with spoligotypes from almost 40.000 *M. tuberculosis* isolates, this 4th version allows for new and robust inferences in phylogeography, population genetics, and global epidemiology of *M. tuberculosis* to be drawn.

Description

The description of SpolDB4, per country of isolation of the clinical isolates, or in very rare cases, by country of origin of the patients, is as follows.

AFR, (n = 3121): Angola (n = 4), Burundi (n = 18), Benin (n = 4), Botswana (n = 1), Burkina Faso (n = 1), Centrafrican republic (n = 154), Ivory Coast (n = 84), Cameroon (n = 498), Congo Democratic (former Zaire) (n = 7), Congo (n = 20), Djibouti (n = 5), Algeria (n = 137), Egypt (n = 60), Eritrea (n = 8), Ethiopia (n = 153), Gabon (n = 2), Ghana (n = 1), Guinea (n = 7), Gambia (n = 1), Guinea-Bissau (n = 217), Kenya (n = 71), Libya (n = 54), Morocco (n = 127), Mali (n = 5), Mozambique (n = 28), Mauritania (n = 4), Malawi (n = 122), Namibia (n = 79), Niger (n = 1), Nigeria (n = 6), Rwanda (n = 8), Sudan (n = 41), Senegal (n = 69), Sierra Leone (n = 3), Somalia (n

= 226), Tunisia (n = 10), Tanzania (n = 9), Uganda (n = 11), South Africa (n = 564), Zambia (n = 43), Zimbabwe (n = 247), East-Africa (not specified, n = 15).

CAM, (n = 1353): Netherlands Antilles (n = 1), Barbados (n = 6), Costa Rica (n = 1) Cuba (n = 239), Commonwealth of Dominica (n = 1), Guadeloupe (n = 232), Honduras (n = 7), Haiti (n = 375), Mexico (n = 386), Martinique (n = 105).

EUR, (n = 13624): Albania (n = 5), Austria (n = 1456), Belgium (n = 517), Bulgaria (n = 2), Bosnia and Herzegovina (n = 3), Czech Republic (n = 393), Germany (n = 574), Denmark (n = 281), Spain (n = 374), Estonia (n = 115), Finland (n = 347), France (n = 2265), United-Kingdom (n = 1523), Greece (n = 1), Hungary (n = 62), Ireland (n = 1559*, mainly *M. bovis*), Italy (n = 934), Sicily (n = 125), Latvia (n = 138), Liechtenstein (n = 1), Lithuania (n = 3), Luxembourg (n = 1), Moldova (n = 1), Macedonia (n = 7), Netherlands (n = 973), Norway (n = 31), Poland (n = 227), Portugal (n = 336), Romania (n = 26), Russia (n = 986), Sweden (n = 328), Swiss (n = 1), Ukraine (n = 4), Yugoslavia (n = 28).

FEA, (n = 2624): China (n = 145), Indonesia (n = 344), Japan (n = 138), Korea (n = 4), Myanmar (n = 20), Mongolia (n = 19), Malaysia (n = 598), Philippines (n = 237), Singapore (n = 4), Thailand (n = 302), Vietnam (n = 789), Far-east-Asia unspecified (n = 24).

MECA, (n = 2639): Afghanistan (n = 3), Armenia (n = 119) Azerbaijan (n = 71), Bangladesh (n = 676), Comoro Islands (n = 14), Georgia (n = 272), India (n = 483), Iran (n = 110), Iraq (n = 5), Israel (n = 15), Kazakhstan (n = 55), Lebanon (n = 2), Sri Lanka (n = 16), Madagascar (n = 395), Mauritius (n = 21), Nepal (n = 6), Pakistan (n = 90), Reunion Island (n = 16), Saudi Arabia (n = 99), Turkey (n = 170), Yemen (n = 1).

NAM, (n = 9153): Canada (n = 266), Greenland (n = 4), USA unspecified (n = 1690), USA (Alabama, n = 3), USA (New-York, n = 5948), USA (Texas, n = 1242).

OCE, (n = 235): Australia (n = 36), New Zealand (n = 151), French Polynesia (n = 2), USA (Hawaii, n = 46).

SAM, (n = 3176): Argentina (n = 1150*, mainly *M. bovis*), Bolivia (n = 4), Brazil (n = 842), Chile (n = 2), Colombia (n = 1), Ecuador (n = 12), French Guiana (n = 375), Guiana (n = 3), Peru (n = 96), Paraguay (n = 6), Suriname (n = 8), Uruguay (n = 5), Venezuela (n = 672).

Data format

All spoligotypes were converted into the octal format within Excel spreadsheets [86]. The database is maintained under an Access® format, whereas a Bionumerics® version is also regularly updated (Applied Maths, Sint-Marteen-Latem, Belgium). An updated mySQL-Java-based version is in development. The Information system automatically attributes the shared-type (ST) number to all the entries that correspond to an identical spoligotype found in two or more individual patient isolates, whereas, the entries occurring only once are considered as orphan.

Combined automatized-expert based classification of spoligotypes

To be classified by SpolNet, spoligotypes must be expressed under the form of binary vectors of 43 bits. Using clade and Principal Genetic Groups (PGG) clustering [87], which was previously established or collected by our laboratory and others [88], computerized rules have been generated to sort spoligotypes into clades. Each computerized rule is a translation of a global visual recognition rule. This rule may be defined using a combination of four criteria: (a) presence of a block of one or many consecutive bits (b) absence of a block of one or many consecutive bits (c) presence of at least one bit in a given bit interval (d) absence of at least one bit in a given bit interval. The computer science translation of these four criteria required the creation of a positive and a negative rule. The positive rule translates the fact that a bit is absent or present in the vector. The negative rule translates the fact that there must be at least one bit present or absent in a given interval. Three values can be attributed for each bit, "n" = present, "o" = absent, "x" = variable. The software generates a text file, which contains all spoligotypes sorted by family and PGG, and a specific file for each rule containing the spoligotypes which fulfilled the selection criteria. Rules are hierarchical, i.e. some rules are smooth and almost each spoligotype fulfils it (ex: T1 rule), whereas others are tight (ex: LAM12-Madrid1), hence a final multilevel classification scheme for each shared-type from the most precise to the less precise sub lineage/lineage label (ex: LAM3/LAM9/T1).

Secondly, an algorithm establishes the hierarchic links from the entry data file. This algorithm is based on a comparison of vector. The model (assumption) relies on an evolution that proceeds by deletion of a unique block of one to *n* consecutive bits on the DR locus. For a given spoligotype, the algorithm finds all the potential offspring spoligotypes. The result is transferred into a file whose format is directly used by the BioLayout software [89]. Building a file for BioLayout looks like a simple task but it turns out to be a tedious and time-consuming task for file with hundreds of spoligotypes. This file cannot be built without software for thousands of spoligotypes, as it is the case for the SpolDB4 database project. The two developed algorithms are exponential, i.e. the time to sort data files and to generate Biolayout files increases with the size of

the entry data file. As an example, a file with 1000 spoligotypes may take approximately 20 minutes to be sorted and 20 minutes to generate the Biolayout file. New spoligotype-recognition rules may be introduced as a dynamic process, either when new genotypes are discovered, to introduce new hypothesis, or to modify pre-existing rules.

Definition of indices

The global distribution of spoligotyping patterns was assessed within and between the eight studied continental regions: Africa, Central America, Europe, Far East Asia, Middle East and Central Asia, North America, Oceania, and South America. A slight modification in definition of C1 and C2 was introduced, i.e. a spreading Index SI ≥ 25 instead of SI ≥ 30 is now required to define a clone as "epidemic". Briefly, the Spreading Index (SI) represents a mean number of occurences of a clone, independent of the setting (total n° of occurrences for this ST divided by the number of geographical areas where it is found). The reader should refer to Figure 2 of [26], or to Table 1 for full definition of C1 and C2.

Availability

SpolDB4 listing [see additional file 1] is available as supplemental material. A dedicated interactive website, SITVIT1 which will allow SpolDB4 (39525 genotypes) to be queried online is available [93]. In its research format SITVIT1 allows one to enter MLVA data and to automatically detect MLVA clusters [90]. The full list of investigators having contributed to SpolDBs since the origin of the project will also appear on the website. SpotClust results were extracted from [91].

Quality control (QC)

During the SpolDB4 project, we faced increased quality control problems due to an increased recruitment rate and had to exclude some datasets. A common problem was the systematic absence of one spacer on some membranes. These data sets were systematically excluded. Most of these problems were linked to manufacturing defaults in the commercialized membranes. All data sets were checked individually and sometimes audited on source results (autoradiography). Most investigators, to check their procedures, completed a QC form. International Guidelines to increase spoligotyping quality are in progress and will be reported elsewhere. As in all databases, we assume a reasonable (2 to 5% maximum) error rate in data points. However, the problem of QC in high throughput genotyping technique and database science technology is an emerging issue [92]. A list of STs whose distribution did not change between SpolDB3 and 4 is available upon request. These genotypes represent: (1) ongoing genotypes not detected because of no follow-up in a given area, (2) potentially extinct genotypes, (3) potential typing artefacts. Similarly as in SpolDB3, where 5 STs had been suppressed, 22 STs in SpolDB4 could be artefacts (ST422, 424, 425, 454, 456, 540, 547, 551, 553, 556, 571, 870, 886, 887, 900, 901, 908, 1270, 1575, 1608, 1625, 1896).

Authors' contributions

All authors contributed to the spoligotyping data contained in the database at various levels, by locally isolating, identifying clinical isolates, preparing DNA and genotyping clinically isolates of MTC and sending their results to the Pasteur Institute in Guadeloupe. Their relative contribution would be to tedious to mention here, however, they are all part of the paper because such population-based studies could not be done without adequate mycobacteriological diagnostics, i.e., isolation of clinical isolates, identification, drug-susceptibility testing, DNA extraction, genotyping and ultimately by sharing of data, hence without adequate reward of the numerous persons and labs involved in such analysis.

KB managed the SpolDB4 Information System, controlled data quality, did part of the supervised analysis, used SpolNet for Bioinformatical classification, compiled data synthesis. CS initiated the SpolDB project, established many of the contacts, recruited many of the investigators with NR, writing e-mails, checked the Information System developments helped by numerous Computer scientists (Philippe Leremon, Christel Delfino, Philippe Abdoul, Georges Valétudie, are warmly acknowledged). CS did part of the mining with KB and wrote the paper. JRD managed the PHRI New-York database and sent his data regularly to CS and KB. LR and WP were the two most "big account" data providers, together with AG. BL produced the maps (P. Waniez, Philcarto®, IRD, France) and contributed to the statistical analysis.

Additional material

Additional file 1

Supplemental Table: *SpolDB4 listing of all STs, binary description, octal description, distribution per country of isolation and/or of origin when available, clade/subclade label. Country names were chosen according to the ISO3166-three-letter format. "UI" = unknown. Clade/subclade label using spoligotyping only should be taken as presumptive or indicative of a likely clade/subclade belonging but may in some case be misleading and requires in most cases further investigations to confirm the identity of a given isolate. In some instances, mixed patterns (unrecognized) did not unambiguously allow spoligotyping classification, hence an ambiguous final label in this table.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2180-6-23-S1.pdf]

Acknowledgements

Dr. K. Brudey performed this work as part of her doctoral thesis. She has been working at the Institut Pasteur de Guadeloupe for the last 4 years. Her research focuses on molecular epidemiology, genotyping methods and databases building of *Mycobacterium tuberculosis* complex.

This paper was written as part of the EU Concerted Action project QLK2-CT-2000-00630. The "Unité de la Tuberculose et des Mycobactéries" is supported by the Réseau International des Instituts Pasteur et Instituts Associés, Institut Pasteur, Paris, France. We also would like to thank many investigators whose published or unpublished datas were included in SpolDB4, and others who sent a limited number of unpublished spoligotypes.

Our particular thanks go to (in alphabetical order):

R. Aga (USA), N. Ahmed (India). M. Behr (Canada), M. L. Boschiroli (France), F. Boulahbal (Algeria), E. Bouza (Spain), M. Cardodo-Oelemann (Brazil), G. Cangelosi (USA), U. Dahle (Norway), D. Cousins (Australia), J. T. Crawford (USA), S. David. (Portugal), O. Dellagostin (Brazil), E. Desmond (USA), R. Diaz (Cuba), J. Douglas (Hawaii, USA), F. Drobniewski (United Kingdom), R. Durmaz (Turkey), G. Engelmann (Germany), D. El-Baghdadi (Morocco), P. Easterbrook (United Kingdom), G. Fadda (Italy), P. Freidlin (Israel), A. Gibson (United Kingdom), W. Githui (Kenya), N. Guessend (Ivory Coast), W. Haas (Germany), G. Källenius (Sweden), T. Koivula (Sweden), A. Kwara (USA), C. Mammina (Italy), M. C. Martins (Brazil), T. McHugh (United Kingdom), T. Matsumoto (Japan), N. Morcillo (Argentina), A. Nastasi (Italy), A. Nehrlich (Germany), D. N'guyen (Canada), G. Orefici (Italy), R. Oohata (Japan), J. W. Pape (Haiti), L. Parsons (USA), T. Quitugua (USA), M. Ridell (Sweden), A. Riveira (Philippines), M. H. Ferez Saad (Brazil), L. Sechi (Sardinia, Italy), M. Shemko (United Kingdom), H. Soini (Finland), Y. Sun (Singapore), P. Supply (France), A. Tada (Japan), H. Takiff (Venezuela), A. Vaughan (New Zealand), J. de Waard (Venezuela), G. Yanling (China), S. Zanetti (Italy), A. van der Zanden (Netherlands), M. Zanini (Brazil), A. Zink (Germany). The authors apologize for any names that may have been omitted unwillingly. Two anonymous reviewers helped to improve the manuscript and are also warmly acknowledged.

References

1. Kaufmann SHE, Schaible UE: **100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus.** *Trends Microbiol* 2005, **13(10):**469-475.
2. Mostowy S, Behr MA: **The origin and evolution of *Mycobacterium tuberculosis*.** *Clin Chest Med* 2005, **26:**207-216.
3. Groenen PMA, Bunschoten AE, vanSoolingen D, vanEmbden JDA: **Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method.** *Mol Microbiol* 1993, **10(5):**1057-1065.
4. Jansen R, van Embden JD, Gaastra W, Schouls LM: **Identification of a novel family of sequence repeats among prokaryotes.** *Genomics* 2002, **6(1):**23-33.
5. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60(2):**174-182.
6. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151(Pt 3):**653-663.
7. Gori A, Bandera A, Marchetti G, Degli Esposti A, Catozzi L, Nardi GP, Gazzola L, Ferrario G, van Embden JD, van Soolingen D, *et al.*: **Spoligotyping and *Mycobacterium tuberculosis*.** *Emerg Infect Dis* 2005, **11(8):**1242-1248.
8. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats.** *Microbiol* 1998, **144:**1189-1196.
9. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome.** *Mol Microbiol* 2000, **36:**762-771.
10. Lindstedt BA: **Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria.** *Electrophoresis* 2005, **26(13):**2567-2582.
11. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM: **Stable association between strains of *Mycobacterium tuberculosis* and their human host populations.** *Proc Natl Acad Sci USA* 2004, **101(14):**4871-4876.
12. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N: ***Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS IS6110, VNTR and DR-based spoligotyping suggests the existence of two new phylogeographical clades.** *J Mol Evol* 1081, **53:**680-689.
13. Supply P, Warren RM, Banuls AL, Lesjean S, Van Der Spuy GD, Lewis LA, Tibayrenc M, Van Helden PD, Locht C: **Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area.** *Mol Microbiol* 2003, **47(2):**529-538.
14. Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O: **Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography.** *Genome Res* 2005, **15(10):**1357-1364.
15. Fabre M, Koeck JL, Le Fleche P, Simon F, Herve V, Vergnaud G, Pourcel C: **High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of *hsp*65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*".** *J Clin Microbiol* 2004, **42(7):**3248-3255.
16. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Destro-Bisol G, Coia V, *et al.*: **A back migration from Asia to sub-Saharan Africa is supported by high- resolution analysis of human *Y-chromosome haplotypes*.** *Am J Hum Genet* 2002, **70(5):**1197-1214.
17. Kinsella RJ, Fitzpatrick DA, Creevey CJ, McInerney JO: **Fatty acid biosynthesis in *Mycobacterium tuberculosis*: lateral gene transfer, adaptive evolution, and gene duplication.** *Proc Natl Acad Sci USA* 2003, **100(18):**10320-10325.
18. Klovdahl AS, Graviss EA, Yaganehdoost A, Ross MW, Wanger A, Adams GJ, Musser JM: **Networks and tuberculosis: an undetected community outbreak involving public places.** *Soc Sci Med* 2001, **52(5):**681-694.
19. Hopcroft J, Khan O, Kulis B, Selman B: **Tracking evolving communities in large linked networks.** *Proc Natl Acad Sci USA* 2004, **101(Suppl 1):**5249-5253. Epub 2004 Feb 5242.
20. Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC: **Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics.** *Ann Rev Ecol Syst* 1987, **18:**489-522.
21. Knowles LL: **The burgeoning field of statistical phylogeography.** *J Evol Biol* 2004, **17(1):**1-10.
22. van Belkum A, Struelens M, deVisser A, Verbrugh H, Tibayrenc M: **Role of genomic typing in Taxonomy, evolutionary genetics, and microbial epidemiology.** *Clin Microbiol Rev* 2001, **14(3):**547-560.
23. Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N: **Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission.** *Emerg Inf Dis* 1999, **5(3):**404-414.
24. Sola C, Filliol I, Guttierez C, Mokrousov I, Vincent V, Rastogi N: **Spoligotype database of *Mycobacterium tuberculosis*: Biogeographical distribution of shared types and epidemiological and phylogenetic perspectives.** *Emerg Inf Dis* 2001, **7:**390-396.
25. Filliol I, Driscoll JR, Van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Anh DD, Barlow R, Banerjee D, Bifani PJ, *et al.*: **Global distribution of *Mycobacterium tuberculosis* spoligotypes.** *Emerge Inf Dis* 2002, **8(11):**1347-1350.
26. Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valetudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, *et al.*: **Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study.** *J Clin Microbiol* 2003, **41(5):**1963-1970.

27. Sebban M, Mokrousov I, Rastogi N, Sola C: **A Data-mining approach to Spacer Oligonucleotide Typing of *Mycobacterium tuberculosis*.** *Bioinformatics* 2002, **18:**235-243.
28. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM: **Single-Nucleotide Polymorphism-Based Population Genetic Analysis of *Mycobacterium tuberculosis* Strains from 4 Geographic Sites.** *J Infect Dis* 2006, **193(1):**121-128.
29. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, Fyfe J, Garcia-Garcia L, Rastogi N, Zozio T, *et al.*: **The Global Phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis : insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems and recommendations for a minimal standard SNP set.** *J Bacteriol* 2006 in press.
30. Warren RM, Streicher EM, Sampson SL, Van Der Spuy GD, Richardson M, Nguyen D, Behr MA, Victor TC, Van Helden PD: **Microevolution of the Direct Repeat Region of *Mycobacterium tuberculosis*: Implications for Interpretation of Spoligotyping Data.** *J Clin Microbiol* 2002, **40:**4457-4465.
31. Quitugua TN, Seaworth BJ, Weis SE, Taylor JP, Gillette JS, Rosas II, Jost KC Jr, Magee DM, Cox RA: **Transmission of drug-resistant tuberculosis in Texas and Mexico.** *J Clin Microbiol* 2002, **40:**2716-2724.
32. vanSoolingen D, vanderZanden AGM, deHaas PEW, Noordhoek GT, Kiers A, Foudraine NA, Portaels F, Kolk AHJ, Kremer K, vanEmbden JDA: **Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers.** *J Clin Microbiol* 1998, **36(7):**1840-1845.
33. Aranaz A, Liebana E, Gomez-Mampaso E, Galan JC, Cousins D, Ortega A, Blazquez J, Baquero F, Mateos A, Suarez G, *et al.*: ***Mycobacterium tuberculosis* subsp. caprae subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain.** *Int J Syst Bacteriol* 1999, **49:**1263-1273.
34. van der Zanden AG, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, Nagelkerke NJ, van Soolingen D: **Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides.** *J Clin Microbiol* 2002, **40(12):**4628-4639.
35. Brudey K, Gutierrez MC, Vincent V, Parsons LM, Salfinger M, Rastogi N, Sola C: ***Mycobacterium africanum* Genotyping Using Novel Spacer Oligonucleotides in the Direct Repeat Locus.** *J Clin Microbiol* 2004, **42(11):**5053-5057.
36. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155(2):**945-959.
37. **Structure** [http://pritch.bsd.uchicago.edu]
38. Vijaya-Bhanu N, van Soolingen D, van Embden JDA, Dar L, Pandey RM, Seth P: **Predominance of a novel *Mycobacterium tuberculosis* genotype in the Delhi region of India.** *Tuberculosis* 2002, **82(2\3):**105-112.
39. Singh UB, Suresh N, Vijaya Bhanu N, Arora J, Pant H, Sinha S, Aggarwal RC, Singh S, Pande JN, Sola C, *et al.*: **Predominant Tuberculosis Spoligotypes, Delhi, India.** *Emerg Infect Dis* 2004, **10(6):**1138-1142.
40. McHugh TD, Batt SL, Shorten RJ, Gosling RD, Uiso L, Gillespie SH: ***Mycobacterium tuberculosis* lineage : a naming of the parts.** *Tuberculosis (Edinb)* 2005, **85:**127-136.
41. Douglas JT, Qian L, Montoya JC, Musser JM, Van Embden JD, Van Soolingen D, Kremer K: **Characterization of the Manila Family of *Mycobacterium tuberculosis*.** *J Clin Microbiol* 2003, **41(6):**2723-2726.
42. Namwat W, Luangsuk P, Palittapongarnpim P: **The genetic diversity of *Mycobacterium tuberculosis* strains in Thailand studied by amplification of DNA segments containing direct repetitive sequences.** *Int J Tuberc Lung Dis* 1998, **2(2):**153-159.
43. Kovalev SY, Kamaev EY, Kravchenko MA, Kurepina NE, Skorniakov SN: **Genetic analysis of *Mycobacterium tuberculosis* strains isolated in Ural region, Russian federation, by MIRU-VNTR genotyping.** *Int J Tuberc Lung Dis* 2005, **9(7):**746-752.
44. Zozio T, Allix C, Gunal S, Saribas Z, Alp A, Durmaz R, Fauville-Dufaux M, Rastogi N, Sola C: **Genotyping of *Mycobacterium tuberculosis* in two cities of Turkey suggests a phylogeographical specificity for the LAM7 lineage.** *BMC Microbiol* 2005, **5(44):**.
45. Niobe-Eyangoh SN, Kuaban C, Sorlin P, Cunin P, Thonnon J, Sola C, Rastogi N, Vincent V, Gutierrez MC: **Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon.** *J Clin Microbiol* 2003, **41(6):**2547-2553.
46. Ngo Niobe-Eyangoh S, Kuaban C, Sorlin P, Cunin P, Thonnon J, Sola C, Rastogi N, Vincent V, Gutierrez MC: **Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon.** *J Clin Microbiol* 2003, **41(6):**2547-2553.
47. Easterbrook PJ, Gibson A, Murad S, Lamprecht D, Ives N, Ferguson A, Lowe O, Mason P, Ndudzo A, Taziwa A, *et al.*: **High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a molecular epidemiological study.** *J Clin Microbiol* 2004, **42(10):**4536-4544.
48. Garcia de Viedma D, Bouza E, Rastogi N, Sola C: **Analysis of *Mycobacterium tuberculosis* genotypes in Madrid : description of two new families specific to Spain-related settings.** *J Clin Microbiol* 2005, **43:**1797-1806.
49. Sola C, Ferdinand S, Mammina C, Nastasi A, Rastogi N: **Genetic Diversity of *Mycobacterium tuberculosis* in Sicily Based on Spoligotyping and Variable Number of Tandem DNA Repeats and Comparison with a Spoligotyping Database for Population-Based Analysis.** *J Clin Microbiol* 2001, **39(4):**1559-1565.
50. Hermans PW, Messadi F, Guebrexhaber H, Soolingen Dv, Haas PEWd, Heersma H, Neeling Hd, Ayoub A, Portaels F, Frommel D, *et al.*: **Analysis of the Population Structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia and the Netherlands: usefulness of DNA Typing for Global Tuberculosis Epidemiology.** *J I D* 1995, **171:**1504-1513.
51. Lari N, Rindi L, Sola C, Bonanni D, Rastogi N, Tortoli E, Garzelli C: **Genetic diversity determined on the basis of *kat*G463 and *gyr*A95 polymorphisms, spoligotying, and IS6110 typing of the *Mycobacterium tuberculosis* complex isolates from Italy.** *J Clin Microbiol* 2005, **43:**1617-1624.
52. Ohata R, Tada A: **[Beijing family and other genotypes of *Mycobacterium tuberculosis* isolates in Okayama district].** *Kekkaku* 2004, **79(2):**47-53.
53. Sola C, Zozio T, Ellermeier C, Sajduda A, Naumann L, Nguyen D, Behr M, de Haas P, vanH est R, van Soolingen D, *et al.*: **The presumed origin of a recent tuberculosis outbreak among the Inuit community of Nunavik.** *26th Annual Congress of the European Society of Mycobacteriology:2005: Istanbul, Turkey, June 25–29th 2005* 2005:91. Abstract Book, P-27 poster
54. Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM: **Genome-Wide Analysis of Synonymous Single Nucleotide Polymorphisms in *Mycobacterium tuberculosis* Complex Organisms. Resolution of genetic relationships among closely related microbial strains.** *Genetics* 2002, **162(4):**1533-1543.
55. Baker L, Brown T, Maiden MC, Drobniewski F: **Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*.** *Emerg Inf Dis* 2004, **10(9):**1568-1577.
56. Dale JW, Al-Ghusein H, Al-Hasmi S, Butcher PD, Dickens A, Drobniewski F, Forbes KJ, Gillespie S, Lamprecht D, McHugh TD, *et al.*: **Evolutionary relationships amongst isolates of *Mycobacterium tuberculosis* with few copies of IS *6110*.** *J Bacteriol* 2003, **185(8):**2555-2562.
57. Warren RM, Victor TC, Streicher EM, Richardson M, van der, Spuy GD, Johnson R, Chihota VN, Locht C, Supply P, van Helden PD: **Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS6110 copy numbers.** *J Clin Microbiol* 2004, **42(12):**5774-5782.
58. Soini H, Pan X, Teeter L, Musser JM, Graviss EA: **Transmission dynamics and molecular characterization of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110* [In Process Citation].** *J Clin Microbiol* 2001, **39(1):**217-221.
59. Berkhin P: **Survey of clustering data mining techniques, Accrue Software,. 2002.** 2002. http://wwwaccruecom/products/rp_cluster_reviewpdf
60. Vitol I, Driscoll J, Kurepina N, Kreiswirth B, Bennett K: **SpotClust : a tool to cluster spoligotype data for tuberculosis evolution and epidemiology.** *Recomb 2005: Cambrdige, Ma, May 14–18; 2005* 2005.

61. Duchene V, Ferdinand S, Filliol I, Guégan JF, Rastogi N, Sola C: **Phylogenetic reconstruction of the *Mycobacterium tuberculosis* complex within four settings of the Caribbean region : tree comparative analysis and first appraisal on their phylogeography.** *Infect Gen Evol* 2004, **4:**5-14.
62. Sola C, Rastogi N: **Genetic description and frequency maps of some major families of *Mycobacterium tuberculosis*.** In *Molecular Epidemiology and Population Genetics of Tuberculosis* Edited by: Ngeow YF, SF Yap. Kuala Lumpur: Academy of Sciences of Malaysia; 2006:23-68.
63. Qian L, Embden JDAv, Zanden AGMvd, Weltevreden EF, Duanmu H, Douglas JT: **Retrospective analysis of the Beijing family of *Mycobacterium tuberculosis* in preserved lung tissues.** *J Clin Microbiol* 1999, **37(2):**471-474.
64. Glynn JR, Whiteley J, Bifani PJ, Kremer K, Van Soolingen D: **Worldwide Occurrence of Beijing/W Strains of *Mycobacterium tuberculosis*: A Systematic Review.** *Emerg Infect Dis* 2002, **8(8):**843-849.
65. Phyu S, Jureen R, Ti T, Dahle UR, Grewal HM: **Heterogeneity of *Mycobacterium tuberculosis* isolates in Yangon, Myanmar.** *J Clin Microbiol* 2003, **41(10):**4907-4908.
66. Kulkarni S, Sola C, Filliol I, Rastogi N, Kadival G: **Spoligotyping of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mumbai, India.** *Res Microbiol* 2005, **156(4):**588-596. Epub 2005 Feb 2007.
67. Gascoyne-Binzi DM, Barlow RE, Essex A, Gelletlie R, Khan MA, Hafiz S, Collyns TA, Frizzell R, Hawkey PM: **Predominant VNTR family of strains of *Mycobacterium tuberculosis* isolated from South Asian patients.** *Int J Tuberc Lung Dis* 2002, **6(6):**492-496.
68. Farnia P, Mohammadi F, Masjedi MR, Varnerot A, Zarifi AZ, Tabatabee J, Douraghei M, Ghazisaeedi K, Mansorri D, Bahadori M, *et al.*: **Evaluation of tuberculosis transmission in Tehran: using RFLP and spoligotyping methods.** *J Infect* 2004, **49(2):**94-101.
69. Kempf MC, Dunlap NE, Lok KH, Benjamin WH Jr, Keenan NB, Kimerling ME: **Long-term molecular analysis of tuberculosis strains in Alabama, a state characterized by a largely indigenous, low-risk population.** *J Clin Microbiol* 2005, **43(2):**870-878.
70. Liens B, Sola C, Brudey K, Rastogi N, and the european co-investigators of the SITVIT consortium: **Spatial Genetics and the spreading history of tuberculosis in Europe.** *26th Annual Congress of the European Society for Mycobacteriology June 26th–29th 2005: Istanbul, Turkey, 26–29 Jun 2005* 2005:65. Abstract Book, P-1 poster
71. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, Supply P, Vincent V: **Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*.** *PLoS Pathog* 2005, **1(1):**e5.
72. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, *et al.*: **A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes.** *Am J Hum Genet* 2002, **70(5):**1197-1214.
73. Mokrousov I, Ly HM, Otten T, Lan NLT, Vyshnevskyi B, Hoffner S, Narvskaia OV: **Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype : clues from human phylogeography.** *Genom Res* 2005, **15(10):**1357-1364.
74. Shamputa IC, Rigouts L, Eyongeta LA, El Aila NA, van Deun A, Salim AH, Portaels F: **Frequency of mixed *M. tuberculosis* strains in pulmonary tuberculosis from a high incidence setting.** *Third meeting of concerted action project : new generation genetic markers and techniques for the epidemiology and control of tuberculosis: 2003; Prague* 2003:22.
75. Feil EJ, Smith JM, Enright MC, Spratt BG: **Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data.** *Genetics* 2000, **154(4):**1439-1450.
76. Gibson A, Brown T, Baker L, Drobniewski F: **Can 15-Locus Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Analysis Provide Insight into the Evolution of *Mycobacterium tuberculosis*?** *Appl Environ Microbiol* 2005, **71(12):**8207-8213.
77. Sun YJ, Bellamy R, Lee AS, Ng ST, Ravindran S, Wong SY, Locht C, Supply P, Paton NI: **Use of Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing To Examine Genetic Diversity of *Mycobacterium tuberculosis* in Singapore.** *J Clin Microbiol* 2004, **42(5):**1986-1993.
78. Blackwood KS, Al-Azem A, Elliott LJ, Hershfield ES, Kabani AM: **Conventional and molecular epidemiology of tuberculosis in Manitoba.** *BMC Infect Dis* 2003, **3(1):**18.
79. Scott AN, Menzies D, Tannenbaum T, Thibert L, Kozak R, Joseph L, Schwartzman K, Behr MA: **Sensitivities and Specificities of Spoligotyping and Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing methods for studying Molecular epidemiology of Tuberculosis.** *J Clin Microbiol* 2005, **43:**89-94.
80. Malik AN, Godfrey-Faussett P: **Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease.** *Lancet Infect Dis* 2005, **5(3):**174-183.
81. Smith NH, Dale J, Inwald J, Palmer S, Gordon SV, Hewinson RG, Maynard Smith J: **The population structure of *Mycobacterium bovis* in Great Britain : clonal expansion. published online before print December 1st, 2003.** *Proc Natl Acad Sci USA* 2003.
82. van Soolingen D, Qian L, de Haas PEW, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JDA: **Predominance of a Single Genotype of *Mycobacterium tuberculosis* in Countries of East Asia.** *J Clin Microbiol* 1995, **33:**3234-3238.
83. **SpolDB4** [http://www.pasteur-guadeloupe.fr/tb/spoldb4]
84. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, Van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, *et al.*: **Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology.** *J Clin Microbiol* 1997, **35:**907-914.
85. van Embden JDA, van Gorkom T, Kremer K, Jansen R, van der Zeijst BAM, Schouls LM: **Genetic variation and evolutionary origin of the Direct repeat locus of *Mycobacterium tuberculosis* complex bacteria.** *J Bacteriol* 2000, **182:**2393-2401.
86. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, Heersma H, Lillebaek T, Quitugua T, Rastogi N, *et al.*: **Spacer oligonucleotide typing of *Mycobacterium tuberculosis* : recommendations for standardized nomenclature.** *Int J Tuberc Lung Dis* 2001, **5:**216-219.
87. Sreevatsan S, Pan X, Stockbauer K, Connell N, Kreiswirth B, Whittam T, Musser J: **Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination.** *Proc Natl Acad Sci USA* 1997, **94:**9869-9874.
88. Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM: **Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping.** *J Clin Microbiol* 2000, **38(2):**669-676.
89. Enright AJ, Ouzounis CA: **BioLayout– an automatic graph layout algorithm for similarity visualization.** *Bioinformatics* 2001, **17(9):**853-854.
90. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.** *J Clin Microbiol* 2001, **39:**3563-3571.
91. **SpotClust** [http://www.rpi.edu/~bennek/EpiResearch]
92. Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, Foote SJ: **Identification and analysis of error types in high-throughput genotyping.** *Am J Hum Genet* 2000, **67(3):**727-736.
93. *Institut Pasteur de la Guadeloupe* [http://www.pasteur-guadeloupe.fr/tb].